建设中文词汇语义资源中的一些问题和我们的对策

董振东　　董强
中科院计算机语言信息工程研究中心
dzd@keenage.com　　DongQinag@keenage.com

摘要：知网是一个中英文双语的描写概念与概念之间的关系以及概念的属性与属性之间的关系的知识系统。自 **1999** 年公布以来，它已在许多方面得到了应用，具有广泛的影响。本文讨论了在建设知网这一语义资源中所遇到的各种理论上和技术上的问题，以及作者对它们的考虑和处理。这些问题包括：词汇语义资源的规模、深度、跨语种、词语的选择、意义的区分和义项的确认、语义描述的策略以及关于意义的计算。
关键词：词汇学；词典学；语义；义原；知网；

Resolutions to Some Problems in Building Chinese Lexical Semantic Resources

Zhendong Dong　　Qiang Dong
Research Centre of Computer & Language Engineering Chinese Academy of Sciences
dzd@keenage.com　　DongQinag@keenage.com

Abstract: HowNet is an on-line common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of the Chinese and their English equivalents. It has gained wide application since it was released in 1999. This paper discusses many issues encountered by the authors in building HowNet and presents the solutions to them. They are size, depth, cross-language, selection of words and phrases, differentiation of meanings, semantic representation of words and phrases, and computing of meanings.
Keywords: lexicology; lexicography; semantics; sememe; HowNet;

首先要明确：这里我们讨论的是面向计算机的词汇语义的研究，或是为计算机而建设的词汇语义资源。本文要介绍的是我们在建设知网中遇到的问题以及我们对这些问题的考虑和处理。不同的面向有不同的任务、不同的问题。它们可以相互借鉴，但不能相互代替。

1. 关于规模

一个能基本满足实用需求的词汇语义库的规模的最低词汇量是多少？我们根据我们曾经进行过的机器翻译系统研发的经验，在建设知网之初就确定其规模不可低于 5 万词语。一个词汇语义库如果低于 2 万词汇，那只能是算一种实验，它既不能满足实用的需要，也不足以证明它自身的理论基础、方法是否正确和可靠。从 2 万到 5 万绝不是简单的量的问题，而是质的问题。

知网知识库的中文词语条目现有 7 万，当然实际上不止，在知网中现在还没有把拼音加上，因此象“花”、“行”、“盛”、“了”、“的”、“重”等在计算时都只算是一个条目。知网知识库中的中文义项总数为 8 万 5 千。知网知识库的英文词语条目现有 7 万，其对应的义项总数为

8 万 6 千。知网知识库的总记录数为 14 万 3 千。

2. 关于深度

这也是与规模有关的。这里所说的深度指的是概念分类体系的深度、概念描述的深度以及词语或概念关系的深度。知网中实体类（一般为名词）、属性类（一般为名词）、以及属性值类（一般为形容词和副词）的分类体系的深度是很浅的，但是其事件类（一般为动词）的分类体系的深度是很深的。由于知网采用的是义原为基的策略，因此分类体系的深度是由义原间的关系决定的。知网中的概念描述的深度以及词语或概念关系的深度基本上是以常识为限。一个重要的原则是注意尽量不要过于深入到某一特定领域的专业知识中去，尽量使之保持平衡。对于面向计算机的资源，多一些专业性词语，并没有坏处，只是不要凑数，不可太滥。像有的资源中的很多词语在 Google 上都查不到，那就不好了。

3. 关于语种

从应用价值看，多语种的词汇语义资源比单一语种的有更多的用途，这是不言而喻的。同时我们还要指出，从工程实施看，也是跨语言的好于单一语种的，虽然工作量大了不少，对语言工作者的要求高了不少。我们认为既然这是一种语义资源，它的架构就应该适应各种不同类型的语言才对。大家知道，知网从一开始就是中英文两种语言同步建设的。从一开始我们就想要考验知网的架构，要证明它能够适应不同的语言。不然的话，它有可能成为一种自我封闭的系统，也就很难得到普遍的认可。在我们建设语义资源时，我们不会不去参考某些已有的资源，包括那些面向人的印刷的资源，但是任何资源都会有它的局限性，完全照搬是不可能的。多语种相互参照，同时并举是好的。但据我们的经验，这里有一条大忌，应切忌让一种语言盲目地跟随着另一种语言。

再一个重要原因是，在区分或拿捏那些差别较微妙的词义的时候，参考另一种语言有时会很有帮助。举一个实际的例子。《现汉》里，“大”有一个义项被描写为“程度深”，译文是“high degree”，例子有“~红”、“~吃一惊”、“天已经~亮”等；另外“深” 有一个义项被描写为“很；十分”，译文是“very；greatly；fully；deeply”，例子有“~恐”、“~表同情”等。我们觉得这个“大”和“深”似乎还不就是“程度深”，并不完全等义于“很”、“十分”。于是我们查英语的“greatly”，其解释中有对应于“great”，而“great”的一个义项是“more than usual”。这样一来，我们找到了一个比“程度”更宽泛的义原。后来我们又在《现汉》关于“高”这一词语下发现下列义项：“在一般标准或平均程度之上”，其例子有“~速度”、“体温~”等。于是我们确立如下的义原体系：

```
|  ├ {StandardValue|标准值}
|  |  ├ {NormalityValue|常态值}
|  |  |  ├ {normal|正常}
|  |  |  ├ {abnormal|不正常}
|  |  |  ├ {GreaterThanNormal|高于正常}
|  |  |  └ {LessThanNormal|低于正常}
```

有了{GreaterThanNormal|高于正常}和{LessThanNormal|低于正常}，对解释“大吃一惊”、“大可不必”以及“牛刀小试”、“小打小闹”中的“大”和“小”，就更妥贴。

4. 关于词语的选择

知网的词语主要选自基于语料库的词表、汉语词典、汉英词典等。我们不追求中文的标准词

表。我们认为可以制定参考词表。与其徒劳地制订标准词表以“断词”，不如教会计算机“合词”的规律。知网在词语选择中所遵循的原则是：是否有利于计算机处理，而不太被语言学的某些说法所束缚。例如知网中收有“接住”和“接到”，这是一般词典或词表不收的。虽然这里的“住”和“到”，都表示“达成”，但这里的“接”这个词却是各自有各自特定的义项。这就是说它们的收录对于歧义的排除有好处。“接”在知网里有 5 个义项，它们分别是：{approach|接近}、{catch|捉住}、{connect|连接}、{meet|会面}、{receive|收受}、{replace|代替}。然而“接住”只可能是{catch|捉住}的达成，在知网中标注为：{catch|捉住:adjunct={Vachieve|达成}}；“接到” 只可能是{ meet|会面}或者{receive|收受}的达成，因此在知网中分别标注为：{meet|会面:adjunct={Vachieve|达成}}和{receive|收受:adjunct={Vachieve|达成}}。

在词语的选择中注意百科知识词语，如主要地名、机构名称、著名的产品品牌等。在注意及时吸收新词语时，要考虑到使用频率、稳定性以及计算机处理的实际。这就是知网不收“打的”、“法轮功”等的原因。

在词语的选择中，可以不收“缩略语”，但必须收它的原语。例如，可以不收“私企”、“婚介”、“军援”等缩略语，但应该收“私营企业”、“私人企业”、“婚姻介绍所”、“军事援助”等等。

5. 关于意义的区分和义项的确认

意义的区分和义项的确认是建设语义资源必然遇到的、往往十分棘手的问题。知网在这个问题上有以下原则。第一，参考现存的词典，但不可完全照搬；第二，遇到难处理的问题时，遵照知网的分类体系的规定。

5.1 事件类（一般为动词）要严格遵循“事件框架”的规定。例如在《现代汉语词典》中，“给”的第一义项是：“使对方得到某些东西或某种遭遇”。而知网中实际上有两个义项对应这一定义，即{give|给}和{CauseToDo|使动}。这是因为它们有着完全不同的上下位关系，如：

event|事件
==> act|行动
==> ActSpecific|实动
==> AlterSpecific|实变
==> AlterRelation|变关系
==> AlterPossession|变领属
==> give|给

event|事件
==> act|行动
==> ActSpecific|实动
==> AlterSpecific|实变
==> AlterState|变状态
==> MakeAct|使之动
==> CauseToDo|使动

{give|给}属于改变领属关系的行为动作，而{CauseToDo|使动}属于改变状态中使受事产生行为动作的行为动作。再者根据知网的“事件关系与角色转换”的描述，{give|给}与{own|有}和{OwnNot|无}是相关的，但{CauseToDo|使动}却完全不同。由此可见，“使对方得到某些东西或某种遭遇”，这个定义对人而言，也许无可非议（但英语词典还是分开两个定义的），但对于计算机就不可取了。到任何时候我们都会提醒自己：人是在有了一定的知识条件下处理语言（学习、理解或表达），但机器是在完全没有知识的条件下处理语言。

又例如，在《现代汉语词典》中，“看”的第一义项是：“使视线接触人或物”，此义项的例子有“~书；~电影”。这就是说《现汉》并不认为中文的“看”有“读”的意义。然而在知网中，“看”既有{look|看}的义项，也有{read|读}这个义项。这是不是被英文所左右呢？不是的。同样是因为在知网中它们有着完全不同的上下位关系，即：

event|事件
==> act|行动
==> ActSpecific|实动
==> AlterSpecific|实变
==> AlterState|变状态
==> AlterMental|变精神
==> AlterKnowledge|变感知
==> MakeOwnKnowledge|使自我感知
==> sense|感觉
==> look|看

event|事件
==> act|行动
==> ActSpecific|实动
==> AlterSpecific|实变
==> AlterState|变状态
==> AlterMental|变精神
==> AlterKnowledge|变感知
==> MakeOwnKnowledge|使自我感知
==> GetKnowledge|认知
==> TryToKnow|弄懂
==> read|读

由此我们可以看到，{look|看}是感知的行为动作，但{ read|读}却是认知的行为动作。“看”的同义词语有“观”、“视”等，其结果可能是“见”。但作为{ read|读}的“看”，它的同义词语有“阅”。

5.2 实体类（一般为名词）要根据“实体分类体系”来定位。这里特别介绍如何处理一个概念可以认为属于两个类的问题。例如“坦克”、“军舰”、“战斗机”等究竟认定是武器呢还是车、船、飞行器呢？由于知网采用的是特征描述而不是简单的分类，因此这样的问题并不难处理。例如“坦克”被标注为{weapon|武器:domain={military|军},restrictive={LandVehicle|车}}。这样一来，它的两个主要特征就都得到体现了，不会影响意义的计算。

5.3 属性值类（一般为形容词或副词），根据我们的体验，这是最难拿捏的一类。例如“壮观”、“宏伟”、“壮丽”等，看上去有区别，它们有着不同的搭配词语，但又好像没有很大的区别，可以是属于同一个义原。在知网中，关于属性和属性值有一条规定，这就是它们两者必须是一一对应的，就是说，有一个属性，就一定会有与之相对应的属性值，同样地，有一个属性值，就一定会有一个与之相对应的属性。我们认为不存在没有属性的属性值，也不存在没有属性值的属性。这样的原则对于拿捏义项有一定的帮助。举一个例子：
“健康”（对应的英文是“healthy”），“健壮” （对应的英文是“sturdy”），它们是不是一类的呢？在知网中它们分属不同的两类。因为前者对应的属性是“健康状况”，而后者对应的属性是“体格”。“健康状况”与“体格”是不同的。身体健康并不一定体格棒。顺便提一句，在 WordNet 中， “healthy”有其对应的属性 – “health”，但是“sturdy”却没有其对应的

属性。

6．关于语义描述的策略

建设一个词汇语义资源，如何描述是首当其冲的问题。我们可以采取义类描述的策略，中文、英文和日文等都有种种印刷的义类词典可供参考。但知网并没有采取义类描述的策略，而是采取特征描述的策略。所谓特征描述有两个要素，第一，提取和确定特征，知网采用的是义原（2200 个）以及可能的关系（110 个）；第二，标注的方法和形式，知网采用的是一套基于义原和关系的结构化的标注语言 — 知识库描述语言（KDML）。试看下面的例子：

NO.=020950
W_C=大学
G_C=N
E_C=
W_E=university
G_E=N
E_E=
DEF={InstitutePlace|场所:domain={education|教育},modifier={HighRank|高等},{study|学习:location={~}},{teach|教:location={~}}}

这里 InstitutePlace|场所、education|教育、HighRank|高等、study|学习、teach|教等均为义原。而 domain、location、:modifier 等均为关系。这一段知识描述语言所表达的意义是直观的，即是“大学”是一个场所，属于教育领域，等级为高等的，在这个场所里人们学习和教书”。再例如：

NO.=020957
W_C=大学生
G_C=N
E_C=
W_E=college student
G_E=N
E_E=
DEF={human|人:{study|学习:agent={~},location={InstitutePlace|场所:domain={education|教育},modifier={HighRank|高等},{study|学习:location={~}},{teach|教:location={~}}}}}

同样地，这一段知识描述语言所表达的意义是直观的，即“大学生”是一个人，他是学习的施事，他在上面描述的场所学习。

由此我们也不难看出，知网的义原的表达以及描述语言不仅直观、可读性较好，最重要的是它们非常方便于意义的计算。 而意义的计算是面向计算机的语义资源的最关键的任务和衡量其质量的最主要的标准。 知网之所以采用基于义原和关系以及描述语言的表达策略的另一个重要原因是考虑到中文词语的特点。下一节将着重说明这一点。

7．关于中文词语的特点

既然是语义描述应该是不依赖特定语言的，即与语言无关的，这是不可否认的。但是如果描述的对象是某一特定语言的词语，那么就有可能这样的描述就不能不反映该特定语言的特点

了。我们认为描述的内容是与具体语言无关的，但是描述的方法有可能是与语言有关的。中文与英文相比，它们的词语常是不对称的。中文里通常被列为所谓的“词”的，它的对译语在英文中可能是一个短语，名词短语、动词短语等。例如：“救国”、“救荒”、“救火”、“救生”、“救灾”、“做客”、“坦陈”、“盛赞”、“严惩”等等。仅仅用一个语义类别来描述它们或仅仅让它们对应到分类体系中的一个简单的上位结点上去，显然是不妥的，因为这样一来将丢失大量的信息。譬如说，假使我们让“救国”、“救荒”、“救火”、“救生”、“救灾”等对应到“救”这个上位结点，显然是不可取的，因为这会大大有损于知识的颗粒度，从而有损于意义的计算。

我们说知网考虑到了中文词语的特性，并不是词语的结构的特性，而是词语的语义组合的特性。采用现有的知网的描述策略，不仅可以从容应付中文，也完全能从从容应付英文。

8．关于意义的计算

把词语视为一种符号，对于这样的符号可以有多层面的不同的计算，如字符串的计算，声音的计算，意义或内容的计算，这就是词语的形、声、义三个层面的计算。在语言技术处理中不同的需求和应用，会用到不同类型的计算。不同类型的计算将用到不同类型的资源。

知网的意义计算的理论基础是：对概念进行基于义原及其关系的、孤立的、静止的描述，并激活概念之间的动态的，相互的关联关系。从上面我们举的“大学”和“大学生”的例子，可以看到我们的做法。我们对于这两个词语的描述是静止的和孤立的，我们并没有描述它们之间有什么关联。但大家一定可以从它们各自的定义中看到它们是通过某些义原而产生特定的关系。再例如，知网并不会描述“医院”和“医生”，或者“医院”和“学校”等有什么关系。但是当我们将它们激活后，它们的关系就会具体地展现在我们的面前。

关于意义的计算，目前有两个热门话题，一是词语的相似度的计算，一是词语的相关性的计算。如今我们正好有了这两方面计算的软件包，它们又都是基于知网的。前者是刘群研发的[8]，后者是董强研发的，称为《相关概念场》。这两个软件包的另一个共同点是它们都是可以任意测试的，而不是仅仅只有几个光鲜的例子而已。词语的相似度和词语的相关性是不同的。用刘群的软件包对“医生”、“医院”、“护士”、“学校”进行相似度测算，我们将得到如下数据：

“医生”和“医院”：0.183563　　“医生”和“护士”：0.948000
“医生”和“学校”：0.140370　　“医院”和“学校”：0.666756
“护士”和“学校”：0.140370

用董强的《相关概念场》软件包对“医生”、“医院”、“护士”、“学校”进行相关性测算，我们得到的结果表明“医生”、“医院”、“护士”三者都在相关场内，但“学校”不在同一的相关场内。其具体数据如下：

“医院”：1231 词语相关，其中包括“医生”和“护士”，但不包括“学校”
“护士”：　647 词语相关，其中包括“医生”和“医院”，但不包括“学校”
“医生”：　590 词语相关，其中包括“医院”和“护士”，但不包括“学校”
“学校”：　546 词语相关，其中不包括“医生”、“护士”、“医院”

《相关概念场》软件包是中英文双语的，我们用英语测试，所得到的数据如下：

“hospital”：1308 词语相关，其中包括“doctor”和“nurse”，但不包括“school”

“nurse”：　706 词语相关，其中包括“doctor”和“hospital”，但不包括“school”
“doctor”：　649 词语相关，其中包括“hospital”和“nurse”，但不包括“school”
“school”：　600 词语相关，其中不包括“doctor”、“nurse”、“hospital”

9．关于人工开发还是自动生成

经常有人问我们，“知网是手工做的还是计算机自动做的”，“为什么不是自动地利用大规模语料来做”，“是不是可以利用 WordNet 和一本英汉词典来自动生成”等等关于建设的方法方面的问题。更有人似乎认为只有自动做出来的才会可靠或才会更好。我们可以见到的情形是：对于全自动生成的大规模语义资源，说的人比做的人多，而做的人比用的人多。我们认为只有那种计算机辅助、基于大规模语料、由人工最后完成的资源，才会是真正可靠的语义资源。

10．结束语

知网已经初具规模，也在多方面得到了应用。但今后我们的工作还很多，主要会在三个方面：第一，开发基于知网的软件包；第二，开拓新的语种；第三，强化英文，争取早日推出可以接近 English HowNet 的资源。届时我们可以对英文方面的应用对 WordNet 和 HowNet 做比较。当然我们也期待着建设 Chinese WordNet 的学者将来有机会也做出上面介绍过的词语相似度软件包和概念相关场软件包，这样可以做一些有益的比较和进一步的讨论，以求共同发展。

引用文献

陈小荷.1998.〈一个面向工程的语义分析体系〉,《语言文字应用》第2期

董振东.1984.〈逻辑语义及其在机译中的应用〉，中国的机器翻译

董振东.1988.〈机器翻译的启示和挑战〉,《上海科技翻译》第1期（总第11期）

Zhendong Dong. 1988. Knowledge description: what, how and who, Manuscripts & Program of International Symposium on Electronic Dictionary, Tokyo

董振东.1998.〈语义关系的表达和知识系统的建造〉,《语言文字应用》第3期

冯志伟.1996.《自然语言的计算机处理》，上海外语教育出版社

贾彦德.1999.《汉语语义学》，北京大学出版社

刘群，李素建.〈基于《知网》的词汇语义相似度的计算〉，第三届汉语词汇语义学研讨会，台北，2002年5月

俞士汶等. 1998.《现代汉语语法信息词典详解》，清华大学出版社

徐通锵. 1998.《马氏文通》和中西语言学结合的道路，’98现代汉语语法学国际学术会议论文提要集

张普. 1998.〈主持人的话〉,《语言文字应用》第2期

在线资源

http://www.keenage.com

http://www..cogsci.princeton.edu/~wn